Lab One Visualization and Data Preprocessing

January 21, 2019

1 Business Understanding (10)

This dataset contains information from schools within the California Department of Education. It specifically addresses school progress on standardized tests in Mathematics during the 2017-2018 school year. In part, the data will be used to support Local Education Agencies (LEA) in identifying strengths, weaknesses, and areas for improvement; assist in determining whether LEAs are eligible for assistance; and assist the SSPI in determining whether LEAs are eligible for more intensive state support/intervention. The school accountability system in California has 10 priorities. This dataset looks at one part of the 4th priority: Student Achievement. The dataset used in this analysis pertains specifically to its Mathematics curriculum and the progress made from 2017 school year and the 2018 school year.
From this dataset, we will look at how the state, each county, LEA, and campus increased or decreased their average distance above or below the passing standard score. This dataset also allows us to look at different programs (English Learner, Socioeconomic Disadvantaged, Students with Disabilities, Foster Youth, Homeless Youth) and race/ethnicities (Black/African American, American Indian or Alaska Native, Asian, Filipino, Hispanic, Pacific Islander, White, Multiples Races/Two or More). This project will address what difference in performance progress were seen between public and charter schools and its respective population subgroups, as described above. The outcome from this will give insight to California on the performance progress in order to help lawmakers better fund schools.

- Is one subgroup out/underperforming another?
- How do districts compare performance wise? County Offices of Education? What about with subgroups?
- Charter v. Traditional schools? What about with subgroups?
- Any impacts/trends by # of kids tested?

Data is important because - It tells you how students performed on the CAASPP (state wide) test - It tells you what achievement/performance gaps exist - so policy can be made to address these issues

2 Data Meaning Type (10)

An is putting this part together

3 Data Quality (15)

At first glance, you will notice that the 'All' category total in any line is less adding up all the lines for similar campuses. This is happening because a student may be in several different categories (i.e., a student may be White (WH), in foster care (FOS), be living in a low socio-economic status (SED) and be served by special education (SWD) – student will 'count' in 4 of the 16 categories listed). No edits were needed for this phenomenon. Additionally, there were columns of data that we choose to eliminate from the data altogether. There were 7 columns at the end of the dataset that give the data for the "California Alternate Assessment". This test is usually given to students served by special education who have profound disabilities and are unable to complete the same test as others. There are approximately 1% of the total population who are included, and the data does not support our current research, therefore, these columns are also deleted. Next, we found a column of data called "pairshare_method". If a campus does not include students of a testing grade level (i.e. PK campuses) or the campus only serves special populations (i.e., mental health hospital), the data is often paired with another campus in the district or charter. If we kept this data in our dataset, these would be duplicates. Therefore, the lines of data were deleted and then the column was deleted. There are many missing values, however, they are not mistakes. The values that are missing appear when a population has 10 or fewer students represented. When this occurs, the data is not calculated for the purpose of this data file and state accountability ratings. As we look at the data, you will notice some of the larger category averages come out to NaN (Not a Number) because each individual total is 10 or fewer. We chose to delete these lines from our dataset because the data we need to calculate is not included for these lines.

Simple Statistics (10) We ran df.describe() on the full dataset to understand the range, average, and quartile of the numerical values. Due to the way the data is setup, this doesn't give us very much information. From this output, however, we can use the 'count' to find out the location of most of our 'NaN'.

I can talk about the stats done with the groupby data here but we may want to run some more 'simple statistics' here so that we have more 'baseline' before we jump to the 'fancy' stuff.

4 Visualize Attributes (15)

Need more 'basic' stuff... the fancy stuff all belongs in the description of the Joint Attributes Maybe pull out a few of the subpops and then take Charter and Public and run some stuff there.?? Just an idea. We need to visualize at least 5 attributes.

5 Explore Joint Attributes (15)

Jodi still needs to add... discuss groupby and cross-tabulation

6 Explore Attributes and Class (10)

Jodi still needs to add

7 New Features (5)

Jodi still needs to add... discuss the combining of fields to create a SchoolType

8 Exceptional Work (10)

Jodi still needs to add…